# **TRUTHREADER:** Towards Trustworthy Document Assistant Chatbot with Reliable Attribution

Dongfang Li<sup>⊕</sup>, Xinshuo Hu<sup>⊕</sup>, Zetian Sun<sup>⊕</sup>, Baotian Hu<sup>⊕</sup>⊠, Shaolin Ye, Zifei Shan, Qian Chen, Min Zhang

<sup>®</sup>Harbin Institute of Technology (Shenzhen), Shenzhen, China {lidongfang,hubaotian,zhangmin2021}@hit.edu.cn

{yanshek.woo, zetiansun.cs, slye0612,zifeishan,qchen.hust}@gmail.com

### Abstract

Document assistant chatbots are empowered with extensive capabilities by Large Language Models (LLMs) and have exhibited significant advancements. However, these systems may suffer from hallucinations that are difficult to verify in the context of given documents. Moreover, despite the emergence of products for document assistants, they either heavily rely on commercial LLM APIs or lack transparency in their technical implementations, leading to expensive usage costs and data privacy concerns. In this work, we introduce a fully open-source document assistant chatbot with reliable attribution, named TRUTHREADER, utilizing adapted conversational retriever and LLMs. Our system enables the LLMs to generate answers with detailed inline citations, which can be attributed to the original document paragraphs, facilitating the verification of the factual consistency of the generated text. To further adapt the generative model, we develop a comprehensive pipeline consisting of data construction and model optimization processes. This pipeline equips the LLMs with the necessary capabilities to generate accurate answers, produce reliable citations, and refuse unanswerable questions. Our codebase, data and models are released at: https://github.com/HITsz-TMG/ TruthReader-document-assistant, and the video demonstration of our system is available at https://youtu.be/RYVt3itzUQM.

#### 1 Introduction

The main objective of the document assistant chatbot is to establish a conversational mode that enables the users to seek relevant information from given documents (Ma et al., 2020; Zhao et al., 2023b). The advent of Large Language Models (LLMs) can greately enhance the capabilities of document assistant chatbots because of their abilities of multilingual understanding, commonsense

|                           | Multi-Docs           | Reference            | Citation   | Attr. Score          | Generator |
|---------------------------|----------------------|----------------------|--|----------------------|-----------|
| Commercial Product        |                      |                      |  |                      |           |
| Three Sigma 1             | <ul> <li></li> </ul> | ~                    | ×  | ×                    | UNK       |
| Aether Brain <sup>2</sup> | ×                    | ~                    | <ul> <li>Image: A second s</li></ul> | ×                    | UNK       |
| ChatPDF 3                 | ×                    | ~                    | <ul> <li>Image: A set of the set of the</li></ul>  | ×                    | UNK       |
| txyz <sup>4</sup>         | ×                    | <ul> <li></li> </ul> | ~  | ×                    | UNK       |
| Open-source Project       |                      |                      |  |                      |           |
| doc-chatbot 5             | ~                    | ×                    | ×  | ×                    | COM       |
| GPT-4 & LangChain 6       | ~                    | ~                    | ×  | ×                    | COM       |
| DocsGPT 7                 | <ul> <li></li> </ul> | ~                    | ×  | ×                    | COM & OS  |
| TRUTHREADER (ours)        | <ul> <li></li> </ul> | <ul> <li></li> </ul> | ~  | <ul> <li></li> </ul> | OS        |

Table 1: Feature comparison between TRUTHREADER and popular commercial (COM) and open-source (OS) document assistants. "UNK" means unkown. "Attr. Score" represents the attribution score.

reasoning, and instruction following (Touvron et al., 2023; OpenAI, 2023). Numerous frameworks and commercial products have emerged that harness LLMs to power their systems as shown in Table 1.

Despite the prosperity of LLM-based document assistants, some critical challenges remain unresolved. On one hand, such products face a high demand for truthfulness, which poses a significant challenge for LLMs, as their inherent generative mechanisms lack explicit factual grounding (Tonmoy et al., 2024). Specifically, LLMs may produce extrinsic hallucinations when essential information is missing from the retrieved documents (Chen et al., 2023b). In this context, (Q1) verifying the factuality of the response is difficult due to the length of background documents and the complexity of the response (Chern et al., 2023; Min et al., 2023; Zhang et al., 2023). On the other hand, (Q2) a common limitation of existing open-source projects is their reliance on commercial APIs. The drawback is manifold: (1) the frameworks using commercial APIs limit the space of optimization on local domains; (2) the technical intricacies of com-

89

<sup>&</sup>lt;sup>™</sup>Corresponding author.

<sup>&</sup>lt;sup>1</sup>https://www.threesigma.ai

<sup>&</sup>lt;sup>2</sup>https://aetherbrain.ai/

<sup>&</sup>lt;sup>3</sup>https://www.chatpdf.com/

<sup>&</sup>lt;sup>4</sup>https://app.txyz.ai/

<sup>&</sup>lt;sup>5</sup>https://github.com/dissorial/doc-chatbot

<sup>&</sup>lt;sup>6</sup>https://github.com/mayooear/

gpt4-pdf-chatbot-langchain

<sup>&</sup>lt;sup>7</sup>https://github.com/arc53/DocsGPT



Figure 1: Screenshot of our document assistant chatbot TRUTHREADER. The left side of the figure displays the interfaces for file uploading and web parsing, along with the corresponding parsed document content. On the right side, the complete document dialogue interface is shown, where questions can be asked in the dialogue window. It is worth noting that the generated responses include *inline citations*, followed by *attribution score*. Clicking on the citation tags allows the attribution window to jump to the corresponding *attribution chunks*.

mercial products are often concealed, impeding further research of the problem within the community. Moreover, (3) the cost to use such products can be high, and the exposure of private documents to commercial APIs raises concerns on data privacy.

To address these challenges, we present our TRUTHREADER, an open-source document assistant chatbot with reliable attribution, towards a transparent and trustworthy system. Our system consists of a conversational document retriever optimized for multi-turn dialogues, and a retrievalaugmented generator to generate answers. (A1) To facilitate the verification of the factual consistency in the generated text, TRUTHREADER enables the LLMs to generate answers with detailed inline citations, which can be attributed to the relevant document chunks (i.e., attribution chunks). Additionally, we incorporate a novel attribution score interface, which measures the consistency between responses and attribution chunks. It enables users to engage in dialogues and enhance the factual grounding of their queries, thereby efficiently reducing hallucination. (A2) Different from the applications that directly utilize commercial LLM APIs, we showcase a pipeline that trains local and controllable retrieval-augmented LLMs from open-source foundation models. Our comprehensive pipeline involves modules for data construction and model optimization, enabling domain adaptation with no requirement on any human-annotated data, making it feasible to adapt to local documents. Overall, our system exhibits the following capabilities: (1) It

excels in generating accurate responses that align with the provided documents; (2) It is capable of identifying and refusing unanswerable questions when inadequate relevant information is available within the documents; (3) Furthermore, it incorporates inline citations, attributing specific chunks of information within the generated responses. With TRUTHREADER, users are able to glean accurate and credible information from the supporting documents, effectively assisting them in informationseeking tasks. We release the code, data and models to facilitate future research and applications.

### 2 User Interface

In this section, we introduce our document assistant chatbot TRUTHREADER illustrated in Figure 1 and elucidate how it interacts with users.

**Document Upload** The document upload feature provides support for uploading files from the local device or inputting webpage URLs for parsing. Users are allowed to upload one or multiple documents<sup>8</sup>, which are accessible on the left side of the interface. Currently, the system offers support for uploaded file formats such as *txt, docx, pdf*, and *markdown*. Once the files are uploaded or webpages are parsed, the documents are segmented into chunks, which are then displayed in the "Document Content" tab below. Users can adjust the

<sup>&</sup>lt;sup>8</sup>Due to limited deployment resources, the maximum number of uploaded documents in the demo system is set to 50, which can be further extended in general.



Figure 2: The architecture and workflow of our document assistant chatbot TRUTHREADER. It consists of three components: **Document Preprocessor, Conversational Document Retriever**, and **Retrieval-augmented Generator**. The first module is only used for preliminary preprocessing, while the latter two together constitute the workflow of real-time conversation.

chunk size using the slider located above. Optical character recognition (OCR) is available for improved PDF parsing. Additionally, users have the choice to opt for pre-summarization of documents using our adapted LLM.

Chat Interface Upon uploading a document, users can engage in multi-turn dialogues by entering questions pertaining to the document through the dialogue box on the right side. The generative model will generate responses with fragment references based on the retrieved document information. In the situation that no relevant answer is found, the model gives refusal as a response and provides an appropriate explanation. The present conversational abilities primarily encompass the following facets: (1) Multi-document Synthesis: This capability enables classification and collation of multiple articles. For example, "Provide recommendations for AI-related news."; (2) Singledocument Summary: It allows for quick acquisition of the primary details of an article. For example, "What are the main contributions of this paper?"; (3) Question Answering: This feature effectively extracts intricate information from articles. For example, "What is the GDP growth rate mentioned in the document?". Users can switch between different generative models to experience varying model performances. Additionally, the generation behaviour can be controlled by customizing the generation hyperparameters provided below.

Attribution Interaction Attribution interaction serves as a means to identify the source information responsible for generating a response. It enables the verification of factual correctness and the acquisition of additional contextual details. The attribution interaction includes the following aspects:

• Display of citation and reference: The generated response in the chat interface incorporates inline citations, denoted as [1][2]. Furthermore, the references for all retrieved document chunks are listed beneath the response.

- **Display of attribution chunks:** The "Attribution Chunks" tab exhibits the content of each retrieved chunk. The chunks contributing to the citations in the current response are highlighted in bold.
- Display of attribution score: In order to evaluate the consistency between the generated response and attribution chunks, an attribution score progress bar is positioned alongside the citation. <sup>9</sup> As the score increases, the progress bar will display various colours, such as red, yellow, and blue.
- Interaction of citations, references, and attribution: By clicking on a citation or reference, the attribution window automatically redirects to the corresponding paragraph. This functionality facilitates cross-checking the attribution text and generated responses, ensuring convenient access to relevant information.

### **3** System Architecture

This section presents the key technical components of our system TRUTHREADER, which together form the entire architecture as shown in Figure 2. The core web application is built on Gradio package (Abid et al., 2019). The detailed model training progress is discussed in §4, encompassing the retriever and generator modules.

**Document Preprocessor** The pre-processing pipeline involves document parsing, segmentation, and embedding. We parse uploaded files individually based on their types using the LangChain (2022) package. For HTML web pages, we manually extract their element contents recursively to preserve the inherent structure of the document. As for PDF OCR, We integrate Nougat model (Blecher et al., 2023) for parsing. Chunk segmentation is performed using line breaks or periods implemented

<sup>&</sup>lt;sup>9</sup>To measure this consistency, we adopt the precision score of ROUGE-1 due to its efficiency, though it can be replaced by any other factual measurement.

in LangChain. These segmented chunks are then embedded into vectors using our conversational document retriever model and stored for retrieval.

**Conversational Document Retriever** We embed the dialogue by concatenating the current round question with the dialogue history to retrieve the relevant document chunks. Our retrieval model, BGE M3 Embedding (Chen et al., 2024), is fine-tuned on our collected multi-turn document retrieval data. We utilize the Faiss library (Douze et al., 2024) as our vector database for embedding storage and similarity search. In this work, we retrieve 4 chunks for response generation in the subsequent stage to balance effectiveness and efficiency.

**Retrieval-augmented Generator** We implement a retrieval-augmented generator that utilizes retrieved document chunks to prompt LLMs to answer questions. The document chunks are sorted in their natural order and labelled numerically such as [1][2]. Our generator module incorporates three independently pretrained LLMs: Mixtral-7Bx2-Chat (Jiang et al., 2024) <sup>10</sup> and Qwen-14B-Chat (Bai et al., 2023), which are further fine-tuned to enhance dialogue capability. Through this fine-tuning process, the LLMs have acquired the capability to generate inline citations directly within their generated responses, thereby facilitating the display of attribution text.

### 4 Implementation

### 4.1 Conversational Document Retriever

**Data Source** Our study incorporates a fine-tuned retrieval embedding model to enhance conversational document retrieval. Specifically, we utilize dialogues and document pairs from both the Chinese and English datasets of RefGPT (Yang et al., 2023). Each dialogue session, comprising multiple rounds of questions and answers, alongside its historical context, is considered as distinct data, resulting in a training dataset of nearly 400k examples. The instruction template of retrieval query is presented in Table 3, where we concatenate the question-answer pairs from the dialogue history to the current question in reverse order.

**Dialogue Augmentation** To handle topic shifts in conversations, we introduce augmentation techniques involving irrelevant dialogues. We employ

embedding similarity to retrieve somewhat related but ultimately irrelevant dialogue histories. These retrieved histories were subsequently concatenated with partial training for augmentation. The augmented dialogue histories consisted of 4 distinct types: (1) no dialogue history; (2) only relevant dialogue history; (3) only irrelevant dialogue history, indicating a topic transition; (4) both irrelevant and relevant dialogue histories, indicating a previous topic transition.

**Retriever Training** For training, we generated offline hard negative data once, and subsequently trained the model by InfoNCE loss (van den Oord et al., 2018) for 1 epoch. The length of both queries and documents is truncated to 512.

#### 4.2 Retrieval-augmented Generator

We introduce our comprehensive pipeline consisting of data construction and model optimization processes, which enhances the capabilities of LLMs to maintain factual consistency, generate reliable citations, and abstain hallucinatory responses.



Figure 3: The pipeline of data construction.

### 4.2.1 Data Construction Pipeline

**Data Collection** To facilitate LLMs with multiskills, we collect training data from the following aspects:

• Multi-document Synthesis Data We adopt the Self-Instruct method (Wang et al., 2023a) to generate lots of diverse instructions from some seed instructions, e.g., recommend some cutting-edge technology news. Then we couple the generated instructions with retrieved documents from WeiXin Web and generate answers through ChatGPT. <sup>11</sup>

<sup>&</sup>lt;sup>10</sup>We use the version of Mixtral-2x7B-Chat from https: //huggingface.co/cloudyu/Mixtral\_13B\_Chat

<sup>&</sup>lt;sup>11</sup>In this work, we specifically employ ChatGPT based on OpenAI's *gpt-3.5-turbo-0613* as resource limitations.

#### Algorithm 1 Citation Construction Process

- 1: Input Reference  $D = \{D_i\}$ , Response S
- 2: Output Cited response R<sub>c</sub> = {s<sub>i</sub>, c<sub>i</sub>}, where s<sub>i</sub> is a independent sentence.
   3: S ← sentence\_splitter(A)
   4: R<sub>c</sub> ← []
   5: for span s<sub>i</sub> ∈ S do
   6: Citation c<sub>i</sub> ← attributing(D, s<sub>i</sub>)
   7: R<sub>c</sub> ← R<sub>c</sub> ∪ {s<sub>i</sub>, c<sub>i</sub>}
   8: end for
   9: Return R<sub>c</sub>
  - **Single-document Summary Data** We manually create some instructions, e.g., summarize this article, and apply the same method to construct data in Multi-document Synthesis Data.
  - Question Answering Data We utilize several open-source datasets in our research, i.e., RefGPT (Yang et al., 2023) and WebCPM (Qin et al., 2023). Moreover, we generate additional data by leveraging ChatGPT on a diverse range of domains, including but not limited to Wikipedia, news articles, and WeiXin Articles. <sup>12</sup> The data generation process followed the methodology described in RefGPT.

Faithful Filtering Our primary emphasis lies in addressing the issue of entity hallucination filtering, which we have identified as the most significant challenge in LLMs. This aspect is crucial for ensuring faithfulness within the generated outputs. Initially, we employ a filtering approach based on the ROUGE-1 precision scores, comparing the golden answer with the input documents. We assume that examples with scores below a predefined threshold are more likely to exhibit severe hallucinations that are not supported by the input documents. In addition, we filter out examples where the generated answer contains hallucinatory entities that are not present in the input documents. For this purpose, we utilize the Spacy library<sup>13</sup> to implement named entity recognition. The statistics details of the training data of our retrieval-augmented generator are shown in Table 5.

**Citation Construction** We engage in postprocessing of the initial training data to enhance the citation generation capacity of the LLMs. This process involves attributing each sentence in responses to original document segments using more powerful LLM such as ChatGPT. The input structure required for ChatGPT is elucidated in Table 4, and the complete procedural framework adheres to Algorithm 1.

**Refusal Construction** To encourage the LLMs to identify and refuse unanswerable questions that lack sufficient relevant information within the provided documents, we enrich the initial training dataset by incorporating unknown question-response pairs. In detail, we opt for a random subset constituting 10% of the Question Answering Data and substitute the original contextual chunks via citation labels, with somewhat related but ultimately irrelevant chunks. Subsequently, ChatGPT is employed to formulate refusal responses coupled with explanations, which may introduce the primary content of the given documents and elucidate why a particular question is deemed unanswerable.

**Dialogue Augmentation** This step is analogous to the process followed in the conversational document retriever. Please refer to §4.1 for detailed information. Given that WebCPM constitutes a single-turn dataset, we augment it by incorporating one to three dialogue sessions.

**Contextual Augmentaion** To enhance the positional robustness of LLMs towards contextual documents (Liu et al., 2023b), we employ perturbationbased augmentation techniques on the contextual documents. Two primary strategies are utilized for augmentation: (1) shuffling the order of all input contextual documents while updating the reference labels in the answers synchronously, and (2) randomly sampling new documents to replace irrelevant ones within the context. This approach encourages the model to better identify the location of relevant information and improves the accuracy of its responses.

#### 4.2.2 Generator Training

To train the LLMs, we fine-tune them using the negative log likelihood loss for a total of 2 epochs under the learning rate of  $1e^{-5}$ . Specifically, the LLMs are optimized using the LoRA method (Hu et al., 2022). Additionally, the maximum model length is standardized to 4096. Our system is orthogonal to the choice of transformer-based decoder-only autoregressive LLMs.

<sup>&</sup>lt;sup>12</sup>Enterprise data is utilized, even though it is also publicly accessible externally. An unofficial description can be found in https://croud.com/en-gb/resources/ an-introduction-to-wechat-official-accounts/

<sup>&</sup>lt;sup>13</sup>https://spacy.io/

| Model                                  | Answer Accuracy | Refusal Recall | Citation Precision | # Citation |
|--|-----------------|----------------|--------------------|------------|
| Claude-3-Opus                          | 82.95           | 98.86          | 53.28              | 4.43       |
| GPT-4                                  | 82.95           | 100.00         | 92.82              | 2.06       |
| Mixtral-7Bx2-Chat (Jiang et al., 2024) | 73.86           | 34.09          | 73.48              | 2.34       |
| Mixtral-7Bx2-Chat (Adapted)            | 77.27           | 67.05          | 76.67              | 4.17       |
| Qwen1.5-14B-Chat (Bai et al., 2023)    | 86.36           | <u>95.45</u>   | -                  | 0.13       |
| Qwen1.5-14B-Chat (Adapted)             | <u>78.41</u>    | 100.00         | 85.00              | 4.09       |

Table 2: Performance of retrieval-augmented generators. The best are **boldfaced** and the second-best are <u>underlined</u>.



Figure 4: The Recall@4 evaluation results of both the baseline embedding model and our fine-tuned model across different dialogue history types. The mention of "irr & rel" indicates the inclusion of both irrelevant and relevant dialogue histories.

### **5** Evaluation

#### 5.1 Conversational Document Retrieval

**Benchmark** To assess the conversational document retrieval performance, we primarily focus on in-distribution evaluation largely due to the limited availability of a specific test dataset within this specific domain. To overcome this constraint, we create our test dataset using RefGPT, ensuring that it excludes questions and documents from the training data. This process yields 1,919 exemplary instances that serve our evaluation purposes. To expand the pool of document candidates, we sample 20,000 documents from the training data.

**Results** The evaluation results, specifically Recall@1-4 scores, are illustrated in Figure 5 in the appendix, clearly indicating a noticeable improvement achieved through the fine-tuning process. Based on Figure 4, it can be observed that the major improvement of the fine-tuned model lies in its enhanced robustness towards irrelevant dialogue histories, which is particularly important in scenarios involving topic transitions.

#### 5.2 Retrieval-augmented Generation

**Benchmark** We develop an out-of-domain benchmark by leveraging three distinct technical documentation from internal company scenarios.

We have collected real users' query histories and filtered them to obtain single-turn questions that were valuable and difficult. Using our retriever, we can retrieve corresponding document chunks and manually annotate their reference answers, resulting in a total of 88 examples. To evaluate the model's capability to refuse unanswerable questions, we employ the same 88 examples by replacing the original answer-containing fragments with new chunks retrieved from different documents, rendering the questions unanswerable.

**Setting** We conducted a model-based qualitative evaluation to assess the faithfulness of LLMs across three dimensions: (1) Answer Accuracy measures whether the response is correct, based on the human-annotated reference answer; (2) Refusal Recall quantifies the ability of LLMs to appropriately decline unanswerable questions; (3) Citation Precision evaluates the accuracy of the citations generated by LLMs. In line with the methodology employed by Gao et al. (2023), we determined citation correctness by assessing whether the cited document entails the sentence in question. Our evaluation employed GPT-4 models <sup>14</sup>, which have demonstrated a high degree of consistency with manual assessments (Liu et al., 2023c).

**Results** From Table 2, it is evident that both Mixtral and Qwen exhibit excellent performance after optimization. However, Qwen model displays a slight decline in answer accuracy, which could be attributed to post-training it on a well-aligned model. Moreover, both models demonstrate a sufficiently high precision in citing relevant information. The performance would be observed and experienced directly within our online system.

## 6 Related Work

**Document Grounded LLMs** Numerous studies have explored the utilization of LLMs for document readers. Prior works have enhanced the un-

 $<sup>^{14}</sup>$ To evaluate these metrics, we specifically employ GPT-4 based on OpenAI's *gpt-4-0613*.

derstanding of documents by employing sophisticated preprocessing methods (Saad-Falcon et al., 2023; Chen et al., 2023a; Nair et al., 2023; Wang et al., 2024), albeit at a substantial cost. Other approaches have focused on document compression, which is primarily suitable for addressing targeted questions related to specific details within the document (Chevalier et al., 2023; Xu et al., 2023; Liu et al., 2023a; Wang et al., 2023b). However, within the realm of LLMs, there exists a paucity of research concerning the crucial matter of faithfulness in document-based dialogue systems.

Trustworthy LLMs The topic of trustworthiness has long been a subject of interest in the field of generative models (Ji et al., 2023; Zhang et al., 2023). Many previous works aimed at enhancing fact consistency have become less applicable with the advent of LLMs (Shuster et al., 2021; Das et al., 2022; Chiesurin et al., 2023). Recently, several studies have emerged focusing on enabling LLMs to refuse to answer unanswerable questions (Zhao et al., 2023a; Cao, 2023). Teaching models to generate citations has proven to be a valuable approach (Nakano et al., 2021; Menick et al., 2022; Li et al., 2023; Asai et al., 2024; Li et al., 2024; Ye et al., 2024; Zhang et al., 2024; Fierro et al., 2024), facilitating factual attribution and verification of generated responses. While some studies concentrate on fine-grained attribution (Hennigen et al., 2023; Slobodkin et al., 2024; Cao and Wang, 2024; Cohen-Wang et al., 2024), we have chosen the sentence-chunk pair level due to its broader applicability and practicality in common document assistance systems. Leveraging the insights from recent works, our system has been developed to address the issue of multi-faceted truthfulness in document reading.

## 7 Conclusion

This work presents a trustworthy document assistant chatbot, TRUTHREADER, that incorporates incline citation generation and attribution chunks display to enhance the verification of answers. Besides, we propose our pipeline for data construction and model optimization to adapt the LLMs for our system. We hope that this work can contribute to the application and research within the domain of trustworthy document assistant chatbot systems.

### Limitations

**Verification Requirement** While the automation of information retrieval is a core aspect of our system, human verification is still necessary to ensure the factual accuracy of the referenced documents. This necessity arises because our approach is heavily reliant on the correctness of the input documents. If the documents are factually incorrect, the system's output will also be compromised. Therefore, a process for filtering and validating input data is crucial, but it currently remains an area that requires further development.

**Model Scale** Compared to existing commercial products or open-source projects that employ LLMs such as GPT-4 and Gemini, our system utilizes smaller-scale LLMs. Consequently, there may be differences in task diversity and performance when compared to these larger models. Considering the delicate balance between performance and resource, we choose to implement an optimization pipeline, distilling knowledge from larger LLMs to smaller ones. Notably enhancing capabilities in citation generation and negation not only optimizes efficiency but also facilitates wider accessibility and applicability within the developer community.

**Multilingual Capability** Additionally, our system has been primarily optimized for the Chinese context, considering our current application requirements. Although the system retains some capabilities in English, its performance in other languages is comparatively limited. We plan to progressively expand the system's language support to include more languages and extend its application scope in the future.

Attribution Method Despite the emergence of novel attribution methods and models, our research focuses on generating inline citations from input documents. This approach aligns with the most prevalent product format and is highly compatible with existing document assistance systems. We aim to explore multi-grained attribution by integrating chunk-level, sentence-level, and phrase-level analyses. Currently, we utilize ROUGE-1 as the attribution score; however, we plan to incorporate more advanced metrics, such as QAFactEval (Fabbri et al., 2022) and SummaC (Laban et al., 2022), in future work.

### **Ethics Statement**

The datasets of RefGPT (Yang et al., 2023) and WebCPM (Qin et al., 2023), and the documents utilized in our data construction, as well as the Mixtral (Jiang et al., 2024) and Qwen (Bai et al., 2023) models, are available for academic research and non-commercial usage. It is imperative to highlight that the responses produced by our system are derived from language models. Despite extensive training and optimization, our system may sporadically generate errors, demonstrate limited precision, or make inappropriate responses. To ensure the highest level of reliability, we vehemently advise against the exclusive reliance on our system's responses for crucial or significant information. Instead, we recommend supplementing our system's output with additional research, consultation with credible sources, or professional expertise within the relevant field.

### Acknowledgements

This work is jointly supported by grants: National Natural Science Foundation of China (No. 62376067), National Natural Science Foundation of China (No. 62406088) and Guangdong Basic and Applied Basic Research Foundation (2023A1515110078).

### References

- Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Y. Zou. 2019. Gradio: Hassle-free sharing and testing of ML models in the wild. *CoRR*, abs/1906.02569.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May* 7-11, 2024. OpenReview.net.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

- Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *CoRR*, abs/2308.13418.
- Lang Cao. 2023. Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism. *CoRR*, abs/2311.01041.
- Shuyang Cao and Lu Wang. 2024. Verifiable generation with subsentence-level fine-grained citations. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15584–15596. Association for Computational Linguistics.
- Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023a. Walking down the memory maze: Beyond context limit through interactive reading. *CoRR*, abs/2310.05029.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023b. Benchmarking large language models in retrieval-augmented generation. *CoRR*, abs/2309.01431.
- I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.
- Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 3829–3846. Association for Computational Linguistics.
- Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering. In *Findings of the Association* for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, pages 947–959. Association for Computational Linguistics.
- Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. *CoRR*, abs/2409.00729.
- Souvik Das, Sougata Saha, and Rohini K. Srihari. 2022. Diving deep into modes of fact hallucinations in dialogue systems. In *Findings of the Association*

for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 684–699. Association for Computational Linguistics.

- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *CoRR*, abs/2401.08281.
- Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, pages 2587–2601. Association for Computational Linguistics.
- Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. In *Proceedings of the* 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 11397–11417. Association for Computational Linguistics.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 6465–6488. Association for Computational Linguistics.
- Lucas Torroba Hennigen, Shannon Shen, Aniruddha Nrusimha, Bernhard Gapp, David A. Sontag, and Yoon Kim. 2023. Towards verifiable text generation with symbolic references. *CoRR*, abs/2311.09188.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.* OpenReview.net.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang,

Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nlibased models for inconsistency detection in summarization. *Trans. Assoc. Comput. Linguistics*, 10:163– 177.

LangChain. 2022. https://www.langchain.com/.

- Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. 2024. Improving attributed text generation of large language models via preference learning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 5079–5101. Association for Computational Linguistics.
- Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *CoRR*, abs/2311.03731.
- Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023a. TCRA-LLM: token compression retrieval augmented large language model for inference cost reduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9796–9810. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-eval: NLG evaluation using gpt-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 2511–2522. Association for Computational Linguistics.
- Longxuan Ma, Wei-Nan Zhang, Mingda Li, and Ting Liu. 2020. A survey of document grounded dialogue systems (DGDS). *CoRR*, abs/2004.13818.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin J. Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *CoRR*, abs/2203.11147.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 12076–12100. Association for Computational Linguistics.

- Inderjeet Nair, Shwetha Somasundaram, Apoorv Saxena, and Koustava Goswami. 2023. Drilling down into the discourse structure with llms for long document question answering. In *Findings of the Association for Computational Linguistics: EMNLP* 2023, Singapore, December 6-10, 2023, pages 14593– 14606. Association for Computational Linguistics.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browserassisted question-answering with human feedback. *CoRR*, abs/2112.09332.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

- Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Webcpm: Interactive web search for chinese long-form question answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 8968–8988. Association for Computational Linguistics.
- Jon Saad-Falcon, Joe Barrow, Alexa F. Siu, Ani Nenkova, Ryan A. Rossi, and Franck Dernoncourt. 2023. Pdftriage: Question answering over long, structured documents. *CoRR*, abs/2309.08872.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings* of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, pages 3784– 3803. Association for Computational Linguistics.
- Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 3309–3344. Association for Computational Linguistics.
- S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *CoRR*, abs/2401.01313.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.
- Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024. Docllm: A layout-aware generative language model for multimodal document understanding. *CoRR*, abs/2401.00908.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. Self-instruct: Aligning language models with self-generated instructions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 13484–13508. Association for Computational Linguistics.
- Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md. Rizwan Parvez, and Graham Neubig. 2023b. Learning to filter context for retrieval-augmented generation. *CoRR*, abs/2311.08377.
- Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. RECOMP: improving retrieval-augmented lms with compression and selective augmentation. *CoRR*, abs/2310.04408.
- Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023. Refgpt: Dialogue generation of gpt, by gpt, and for GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2535. Association for Computational Linguistics.
- Xi Ye, Ruoxi Sun, Sercan Ö. Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 6237–6251. Association for Computational Linguistics.
- Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme, and Daniel Khashabi. 2024. Verifiable by design: Aligning language models to quote from pretraining data. *CoRR*, abs/2404.03862.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.
- Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, and Jianshu Chen. 2023a. Thrust: Adaptively propels large language models with external knowledge. In *Advances in Neural Information*

Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Yingxiu Zhao, Bowen Yu, Bowen Li, Haiyang Yu, Jinyang Li, Chao Wang, Fei Huang, Yongbin Li, and Nevin L. Zhang. 2023b. Causal document-grounded dialogue pre-training. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, pages 7160–7174. Association for Computational Linguistics.

### **A** Instruction Templates

```
# QUESTION: {{ question<sub>i</sub> }} </s>
# HISTORY:
A: {{ question<sub>i-1</sub> }}
B: {{ answer<sub>i-1</sub> }}
A: {{ question<sub>i-2</sub> }}
B: {{ answer<sub>i-2</sub> }}
```



```
Please add citations to the input text using the given
documents. Citation format: "Text to be cited[1]." or
"Text to be cited[1][2]."
# Demonstration 1
...
# Demonstration 2
...
# Current
Document[1]: {{title1}}{{context1}}
...
Document[n]: {{title1}}{{contextn}}
INPUT: {{answer_snippet}}
OUTPUT:
```

Table 4: The instruction template for ChatGPT to construct citation of our generator data.

We list Table 3 as the instruction template of the retrieval query and Table 4 as the instruction template to construct citations of our generator data.

#### **B** Additional Evaluation Results

As shown in Figure 5, we conducted a Recall@n assessment to measure retrieval performance with and without the incorporation of dialogue history. The results indicate a nuanced impact of dialogue history on the baseline model's effectiveness. Specifically, the baseline model achieved Recall@1 and Recall@4 scores of 58.7 and 69.5, respectively, when dialogue history was excluded, and scores of 58.2 and 70.0 when history was included. This marginal improvement underscores the potential



Figure 5: The Recall @n evaluation results of both the baseline embedding model and our fine-tuned model. "w/o" or "w/" history indicates whether the dialogue history is concatenated with the question.

benefits of context integration. However, the finetuned model with dialogue history demonstrated a pronounced enhancement in performance, achieving Recall@1 and Recall@4 scores of 82.8 and 90.4, respectively. This significant uplift suggests that fine-tuning effectively leverages contextual information, thereby facilitating superior retrieval accuracy. These findings highlight the importance of model adaptation and context utilization in improving the performance of retrieval systems. As shown in Table 6, Qwen generally outperforms Mixtral in terms of macro precision and maintains a high and consistent refuse rate. Mixtral shows variability in its metrics, with notable improvements in citation numbers but a decline in answer accuracy and an increasing refuse rate over epochs. This suggests Qwen may be more reliable in maintaining performance across different metrics, while Mixtral's performance is more variable.

#### C More Details about Datasets

Table 5 provides an overview of the training data utilized for our retrieval-augmented generator, encompassing a variety of sources and languages. The dataset is categorized into five distinct types, each contributing to the robustness and versatility of the model.

**Multi-document Synthesis**: This dataset, in Chinese (zh), comprises 387 examples sourced from WeiXin Subscription Accounts, with answers generated by ChatGPT. This type is crucial for tasks requiring synthesis across multiple documents, enhancing the model's ability to integrate and reconcile information from diverse texts.

**Single-document Summary**: In both Chinese (zh) and English (en), this dataset includes 561 examples derived from WeiXin Subscription Accounts and Wikipedia, summarized by ChatGPT.

| Data                     | Language | Document Source            | Answer Source | #Example |
|--------------------------|----------|----------------------------|---------------|----------|
| Multi-document Synthesis | zh       | WeiXin Articles            | ChatGPT       | 387      |
| Single-document Summary  | zh, en   | WeiXin Articles, Wikipedia | ChatGPT       | 561      |
| QA Created               | zh       | Multi-domains              | ChatGPT       | 1,482    |
| WebCPM                   | zh       | Web                        | Human         | 897      |
| RefGPT                   | zh, en   | Baidu Baike, Wikipedia     | GPT4          | 3,708    |

Table 5: The training data statistics of our retrieval-augmented generator.

| Model              | Answer Acc. | Refusal Rec. | Citation Pre. | # Citation |
|--------------------|-------------|--------------|---------------|------------|
| Mixtral (1 epochs) | 77.27       | 62.50        | 68.35         | 0.93       |
| Mixtral (2 epochs) | 77.27       | 67.05        | 76.67         | 4.17       |
| Mixtral (3 epochs) | 75.00       | 71.59        | 71.65         | 4.96       |
| Qwen (1 epochs)    | 73.86       | 80.01        | 84.31         | 3.69       |
| Qwen (2 epochs)    | 78.40       | 100.0        | 85.00         | 4.09       |
| Qwen (3 epochs)    | 76.13       | 100.0        | 80.12         | 6.14       |

Table 6: Performance of adapted Mixtral-7Bx2-Chat and Qwen-14B-Chat models across different epochs.

This subset focuses on summarization tasks, improving the model's proficiency in condensing information from individual documents.

**QA Created**: Featuring 1, 482 examples in Chinese (zh), this dataset spans multiple domains with answers generated by ChatGPT. It supports the development of the model's capability to handle domain-specific queries, enriching its contextual understanding and response accuracy.

**WebCPM**: Comprising 897 examples in Chinese (zh), sourced from the web and answered by humans, this dataset offers a diverse array of webbased content. It contributes to the model's general knowledge and ability to process and respond to varied web-sourced information.

**RefGPT**: This dataset contains 3, 708 examples in both Chinese (zh) and English (en) from Baidu Baike and Wikipedia, with answers generated by GPT-4. It is instrumental in enhancing the model's ability to reference and utilize structured knowledge from authoritative sources. This dataset broadens the model's linguistic and contextual range, enabling it to handle Chinese and English queries.

The diverse composition of these datasets, including multi-document synthesis, singledocument summarization, domain-specific QA, and reference-based QA in both Chinese and English, equips our retrieval-augmented generator with comprehensive training. This diverse dataset ensures the model's robustness in generating accurate, contextually relevant responses across various types of documents and queries.

### **D** Meta Evaluation

In order to enhance the credibility of our experiments, a meta-evaluation of the automated evaluation method for GPT-4 has been conducted. We primarily evaluated the alignment of GPT-4's accuracy judgments on model-generated answers with human judgments, focusing on a curated test set. Three distinct models were extracted from the development process, and a total of  $264(3 \times 88)$  data points were generated in response to this test set. Subsequently, two domain experts were employed to annotate the accuracy of these model-generated responses. The annotators made judgments based on the given document passages and the standard answers in the test set. Likewise, we also evaluated the annotations provided by GPT-4 for the modelgenerated results. The correlation between human and GPT-4 annotations was calculated, resulting in a Pearson Correlation coefficient of 0.631 and a Spearman Correlation coefficient of 0.631. As a considerable agreement, we conclude that GPT-4 has the ability to effectively replace human evaluation of model-generated results, leading to substantial reductions in costs and time requirements.